Creating an Ontology for Clinical Findings

Mia Levy

Final Report

April 26, 2002

Preceptors

Olivier Bodenreider

Tom Rindflesch

Introduction

Strict knowledge representation criteria stress the importance of distinguishing between the kinds of things that exist in the world and the roles that they play. In the medical domain, clinical *Findings* are categorized by how they are created and by the role they play in the diagnostic process. However, the essence of *Findings* is not well represented in medical ontologies. (Figure 1) The objective of this study is to create an ontology for clinical *Findings* by distinguishing between their Sortal Type and Role, as well as making explicit their relationships to other types. Research is primarily being conducted on the UMLS Metathesaurus and Semantic Network.

Definitions of *Findings* in medical dictionaries and clinically oriented resources such as physical diagnosis texts focus on how *Findings* are created and how they are used. The UMLS Semantic Network defines the semantic type of a *Finding* by how it is created: "that which is discovered by direct observation or measurement of an organism attribute or condition, including the clinical history of the patient." (reference) From this definition we see that *Findings* are created through the process of some healthcare *Activity* and measure some *Attribute*. The *Activities* that create *Findings* include the patient interview, physical examination, laboratory examination and special anatomic and physiologic examinations. It is often the case that the same *Finding* results from more than one *Activity*. For example, the *Finding* "Enlarged Liver" could be the result of palpation and percussion of the abdomen or via abdominal CT scan. Similarly, the *Finding* "Left Ventricular Hypertrophy" can be the result of either an electrocardiogram or an echocardiogram. In this way, *Findings* are not dependent on the *Activity* that produced them for their essence, however some *Activity* must take place in order for a *Finding* to be created.

A *Finding* is also a measure of some *Attribute* under study. This *Attribute* may be anatomic, physiologic, serologic, cognitive, etc. Information science approaches to *Findings* describe them as *Attributes* with either implied value (pain, dyspnea) or with expressed value (temperature of 102°F, increased potassium level) (Sneiderman). With an appropriate set of values a single *Attribute* may produce many *Findings*. Take for example the permutations created by combining the *Attribute* "Cough reflex" with a set of modifiers "impaired, absent, present". The Read Coding system as interpreted by the UMLS creates the hierarchical tree of these permutations as follows:

- Observation of cough (*Activity*)
    - Cough Reflex (*Attribute*)
        - Cough Reflex impaired (*Finding*)
        - Cough Reflex absent (*Finding*)
        - Cough Reflex present (*Finding*)

This hierarchical representation of *Attributes* subsumed under *Activities* and *Findings* subsumed under *Attributes* is a method of representation used by a number of medical vocabularies in the UMLS. (Read Codes, AI/RHEUM, MedDRA) However, the necessity of an *Activity* to measure the *Attribute* to create a *Finding* makes them distinctly different from *Attributes* and makes this hierarchical arrangement inappropriate. While the essence of *Findings* cannot be explained as some type of modified *Attribute*, the relationship of *Findings* to their respective *Attributes* is important in the creation of a well-structured ontology for Findings.

Findings are not only categorized by the way they are created, they are also categorized by the Role they play in the diagnostic process. Clinical *Findings* are used in the medical decision making process to direct diagnostic activities and arrive at a diagnosis. Diseases are patterns of clinical findings connected by a common etiology. (DeGowin) Clinical *Findings* are the catalyst for the differential and definitive diagnosis. *Findings* are often referred to as clues when part of the medical decision making process. Through the cycle of hypothesis making and hypothesis testing *Findings* are progressively transformed into a Diagnosis. (Weed) *Findings* as Roles are commonly classified according to the manner in which they were created. Thus the *Activity* of history taking produces clues called *Symptoms* that are abnormalities perceived by the patient's own senses and often referred to as subjective *Findings*. The clues from the physical examination are called *Signs* where the abnormality is perceived by the examiner's senses. Similarly Laboratory Tests produce *Laboratory Findings*. These clues are used as evidence for disease and drive the diagnostic process through its iterations towards a diagnosis.

As demonstrated above, the relationships for how *Findings* are created and used are well established, but what is their essence? Medical Semiotics, the science of signs, offers some insight by categorizing Findings as signs. This is distinct from the clinical notion of *Signs* described above as clues derived from the physical examination. In Semiotics a sign is a symbol that stands for something to someone. (Peirce) The interpretation of these signs allows us to make sense of the world and thus to diagnose patients. However, the sign and its meaning are usually not the same. Meaning has to be inferred. (Burnum p. 939) The Role that signs play as the agents of inference must be distinguished from the essence of the sign itself.

Further clarification as to the essence of the sign itself is offered by the Data, Information, and Knowledge continuum as defined by the Handbook of Medical Informatics.

- Data is the representation of observations or concepts suitable for communication, interpretation, and processing by humans or machines.
- Information is the meaningful and useful facts extracted from data, or interpreted data.
- Knowledge is the facts and relationships used or needed to obtain insight or to solve problems.

The act of auscultating heart sounds with a stethoscope can be used as an example of the data, information, and knowledge continuum. The sound waves that hit the tympanic membrane are transmitted to the human brain for interpretation. The sound waves represent data in this model. The brain recognizes a pattern in this signal and calls it a heart murmur, elevating the data to the level of information through interpretation. The Role that this information plays in the diagnosis of aortic insufficiency creates knowledge. Similarly, an ECG used to analyze the heart produces voltage signals that are translated to a series of lines on a piece of paper. This data can be interpreted through the analysis of patterns in these lines, such as measurement of the PR interval. The calculated PR interval is compared to a know reference to derive knowledge in the context of disease. Two levels of interpretation exist in this continuum. The first is the cognitive interpretation of data in the creation of Findings. The second is interpretation of the Findings in the context of medical knowledge for use in the diagnostic process. This suggests that Findings may be categorized as Information. This would provide the sortal type for Findings.

Given the central role of Findings in the diagnostic process, it is important to correctly represent them in a medical ontology. Clinical Findings are found throughout electronic medical documents including journal articles and patient records. Knowledge processing of these resources requires a sound representation of the medical domain. Structured ontology principles have been developed for categorizing concepts within ontologies. Strict knowledge representation criteria stress the importance of distinguishing between the kinds of things that exist in the world and the roles that they play. Thus it is important to define the essence of Findings in addition to how they are created and used.

Criteria have been developed that make ex*plicit the distinction between e*ssence *also called Sortal Type and R*ole. Pustejovsky's qualilia structure provides a representation of the defining attributes of an object. Pustejovsky defines four qualia:

- Constitutive: the relation between an object and its constituent parts
- Formal: that which distinguishes it within a larger domain
- Telic: its purpose and function
- Agentive: Factors involved in it origin or bringing it about

Taxonomies of concepts (is-a hierarchies) build upon the formal qualia (what x is). Meronomies (part-of hierarchies) are structures based on the constitutive qualia. The telic qualia (role of x) on the other hand can be represented by either functional categories or associative relationships. (Burgun) Functional categories are hierarchies of

concepts all with Role types. Associative relationships on the other hand more explicitly represent a Role as it relates to another concept.

Guarino and Welty further promote the distinction between Types and Roles by applying the fundamental philosophical notions of Identity, Rigidity and Dependence: (Guarino) (1) Identity is the property of carrying an identity condition (ID), which is a condition that is both necessary and sufficient for study. (2) A property P is Rigid if, for each x, if P(x) is true in one possible world, then it is also true in all possible worlds. The result of diagnostic test is a Rigid property whether or not it is used to diagnose a disease. For example, a body temperature of 98.6°F is a Rigid since the property cannot be lost without losing its Identity. The Role of this concept in the diagnosis of disease however is distinct from its rigid essence. We can imagine this normal physiologic value moving in and out of the Role of clue in the diagnostic process given a specific instance of disease. While the essence of the value is Rigid in all possible worlds, its Role is not. (3) A property P is Dependent if, necessarily, whenever P(x) holds, then Q(y) holds, with $x \neq y$. For example, a body temperature of 103°F is significant in the diagnosis of a disease only if the pattern of *Findings* that make up the disease includes elevated body temperature. *Findings* in this sense are Dependent upon the diseases for their Role as clues.

Guarino and Welty use these fundamental properties to explicitly define Types and Role. Types are Rigid and carry an Identity Condition. Roles on the other hand are Anti-rigid and always Dependent. They are Anti-rigid to ensure that each entity carrying the property does not necessarily carry it and therefore Roles cannot subsume rigid Sortals. A classic example is the distinction between PERSON and STUDENT. The concept PERSON is normally thought of as being Rigid while we can imagine an entity moving in and out of the STUDENT property, Anti-rigid. Further the property STUDENT is Dependent on the existence of a TEACHER and thus fulfills both requirements for being categorized as a Role.

Roles can be further divided into Material Role and Formal Role with specific rules for their hierarchy within an ontology. Both types of Roles are Anti-rigid and Dependent but Material Roles also carry Identity because they inherit the property by subsuming a Type. Formal Roles on the other hand cannot have Identity, and thus cannot be subsumed by Types. They are therefore top-level properties in Role taxonomies also called Functional categories. These ontology principles provide a set of properties and rules for developing a well-structured ontology for Findings.

Materials

Research is primarily being conducted on the National Library of Medicine's UMLS Metathesaurus and Semantic Network. The UMLS Metathesaurus contains over 60 biomedical vocabularies and links together the different names for the same concept. The Semantic Network contains information about the categories to which all Metathesaurus concepts are mapped as well as the permissible relationships between categories. Of the 776940 concepts in the 2002-AA Metathesaurus, 56214 are

categorized as Finding, 11348 as Sign or Symptom, and 4582 as Laboratory or Test Result for a total of 69144 concepts under the parent category of Finding. Since terms within a concept vary syntactically, the research base was extended to include both the preferred form of the concept and its synonyms for a total of 94182 terms. Analysis is also being conducted on the UMLS Semantic Network with its 134 semantic types, 34 semantic relationships and approximately 12000 triplets (type_relationship_type) with special attention to the semantic type Finding and its semantic relationships.

A number of inconsistencies in categorization of concepts have been observed. Some concepts categorized as Findings really belong in the Laboratory or Test Result category. "Abdomen X ray abnormal" (C0740668) for example, is categorized as a Finding when it should be categorized as a Laboratory or Test Result. Some concepts categorized as Findings are not actually Findings but Attributes. For example, the concept "Head circumference" (C0262499) is categorized as both a Finding and an Organism attribute but should only be an Organism attribute since no value modifies the attribute. A number of other concepts categorized as Findings have multiple semantic types, some correct, some incorrect. 4497 of the 53214 concepts categorized as Findings have at least one other semantic type. Of these 4497, 929 are also categorized as Disease or Syndrome, 159 as Organism Attribute.

Further, some concepts that actually are Findings are not categorized as such. For example, the concepts "Ciliospinal reflex positive" (C0558751) and "Cough reflex present" (C0577921) have the semantic type Organ or Tissue Function. Similarly, "Antistreptolysin abnormal" (C0860717) and "Coagulation factor VII level abnormal NOS" (C0855406) are categorized as Laboratory Procedure not as Laboratory or Test Result. A Testing set and a Training set of 1000 terms each were randomly extracted from the 94182 terms. Gold standards created for each set demonstrate that only 84% of the concepts taken from the Finding category were correctly categorized as Findings.

Methods

Two approaches are being utilized to analyze the *Findings*, a top-down approach that is principle driven and a bottom-up approach that is data driven.

Top Down Approach

The top-down approach focuses on determining the ideal Sortal Type for the *Findings*, its hierarchy in the semantic network and the relationships to other semantic types. Methods of analysis include medical dictionary definitions for *Findings* as well as clinically oriented medical texts and journals that discuss the creation of *Findings* and their use in the diagnostic process. The ontology principles for defining Types and Roles as described by Pustejovsky's qualilia and Guarino and Welty were also applied. Existing relationships in the Semantic Network were utilized to place *Findings* in their relational context.

Bottom Up Approach

Methods

Two approaches are being utilized to create an ontology for Findings, a principle driven top-down approach to create the high level structure and a data driven bottom-up approach to populate the ontology.

Top-Down Approach

The top-down approach focuses on determining the ideal sortal type for the Findings, its hierarchy in the semantic network and the relationships to other semantic types. This is accomplished by first creating a formal definition of Findings. The next step is to place within the context of the Data, Information, Knowledge continuum. After creating a formal definition, ontology principles focusing on the distinction between sortal type and roles were be applied to create an ontology for Clinical Findings. Specifically, the principles of Guarino and Welty as well as that of Pustejovsky were used as a foundation for the ontologic structure.

Bottom-Up Approach

The ultimate goal of the bottom up approach is to identify the subcategories for Findings and populate those subcategories with UMLS concepts. Since some of the Findings within the UMLS are not categorized correctly, rules need to be created to extract the true Findings. The primary tool towards both ends is a modified version of the Natural Language Processing program FindX that was originally designed to extract Findings from Medical Records and then applied to Medline Abstracts. (Sneiderman) FindX is build upon tools built by the National Library of Medicine that first perform a syntactic analysis, then map terms to UMLS concepts and finally apply rules for identifying Findings. The rules for identifying Findings are based on the idea that Findings are some kind Activity or Attribute that are modified by a value. Take for example the phrase “Abnormal Liver Biopsy”. A “Biopsy” is a diagnostic activity that is modified by the adjective “Abnormal”. While the biopsy procedure itself is not abnormal, the statement implies that the result of the biopsy procedure is abnormal.

The exact mechanics of the FindX program are as follows. The program first does a syntactic analysis of the input. Take for example the phrase “potassium level increased”. This phrase, like many in the UMLS Metathesaurus and in clinical practice is not in well-formed English. The correct form would be “increase in potassium level. However, the incorrect form is such a pervasive format for Findings that we will use it as an example here. The first step in processing the phrase is a syntactic analysis that uses the Lexical look-up, Xerox Tagger, and Parser programs. The Lexical look-up program, based on the UMLS SPECIALIST Lexicon, finds the parts of speech for each word in the phrase as in the following example:

- potassium - [noun]
- level - [noun, verb, adjective, adverb]
- increased - [verb, adjective]

The Xerox Tagger then resolves the part-of-speech ambiguity.

- potassium - [noun]
- level - [noun]
- increased - [adjective]

The final step in the syntactic analysis is the Parser that defines the head noun and modifiers to create a noun phrase.

- mod (potassium)
- head (level)
- mod (increased)

The noun phrase becomes the input for MetaMap, a program that maps terms to UMLS concepts. MetaMap is highly configurable. For the purposes of the FindX program, MetaMap is configured to the "prefer multiple concepts" setting. Rather than searching the UMLS for the noun phrase as a single unit, the configuration allows components of the phrase to be mapped separately. Once the concepts have been identified, they are then mapped to the UMLS Semantic Network to give their Semantic Type.

- potassium
    - Laboratory Procedure
    - Biologically Active Substance
    - Element, Ion, or Isotope
- level - [Spatial Concept]
- increased - [Functional Concept]

After this preliminary processing, FindX further processes the MetaMap output and the original phrase to determine if it is a clinical Finding. It does this through three stages. First, FindxX groups the UMLS semantic types from the MetaMap output to one of four groupings hand tailored for Findings. For the above example, the Semantic Type Laboratory Procedure maps to the testres grouping in FindX. The testres group represents an Attribute category in the FindX program. FindX then looks through the original phrase and the MetaMap terms for modifiers that match a list of SNOMED modifiers. The modifier "increased" in the example would be marked as a valid value. Similarly it looks for numeric components of the phrase and marks them as values. Finally, FindX applies four rules that determine if a phrase is a Finding. These rules generally look for an Attribute and a value in the phrase. The specific semantic types and values for each rule are described in detail below.

The Anatomy rule is formulated to identify those findings that constitute a comment on some characteristic of an anatomical entity.

- Attribute: UMLS semantic types: 'Acquired Abnormality', 'Body Location or Region', 'Body Part, Organ or Organ Component', 'Body Space or Junction', 'Body System', 'Congenital Abnormality', 'Embryonic Structure', 'Fully Formed Anatomical Structure', 'Tissue', 'Cell', 'Cell Component'
- Value: SNOMED adjective

The Physiologic Function Rule

- Attribute: UMLS semantic types: ‘Physiologic Function’, ‘Organism Function’, ‘Organ or Tissue Function’.
- Value: numeric or SNOMED adjective.

The Test Result Rule identifies findings that are results of diagnostic tests. It was modified slightly from its original form to accommodate Laboratory or Test Results themselves not needing to have modifiers. It was divided into two rules

Test Result 1

- Attribute: UMLS semantic types: ‘Diagnostic Procedure’, ‘Laboratory Procedure’, ‘
- Value: numeric or SNOMED adjective.

Test Result 2

- Laboratory or Test Result’.
- Value: none

The Sign or Symptom Rule

- Attribute: UMLS semantic types: ‘Finding’, ‘Pathologic Function’, ‘Sign or Symptom’
- Value: No value specified.

After evaluating FindX performance in being able to identify UMLS Findings, six new rules were created to increase the categories of Findings that FindX would identify.

Biological Substance Rule:

- Attribute: UMLS semantic types: ‘Amino Acid, Peptide, or Protein’, ‘Carbohydrate’, ‘Lipid’, ‘Nucleic Acid, Nucleoside, or Nucleotide’, ‘Organophosphorus Compound’, ‘Eicosanoid’, ‘Steroid’, ‘Organic Chemical’, ‘Biologically Active Substance’, ‘Hormone’, ‘Neuroreactive Substance or Biogenic Amine’, ‘Enzyme’, ‘Vitamin’, ‘Immunologic Factor’, ‘Receptor’, ‘Substance’, ‘Chemical’, ‘Body Substance’, ‘Organic Chemical’, ‘Element, Ion, or Isotope’, ‘Inorganic Chemical’
- Value: numeric or SNOMED adjective.

Pathogen Rule:

- Attribute: UMLS semantic types: ‘Organism, ‘Bacterium’, ‘Archaeon’, ‘Fungus’, ‘Plant’, ‘Virus’, ‘Rickettsia or Chlamydia’
- Value: numeric or SNOMED adjective.

Therapy Rule

- Attribute: UMLS semantic types: ‘Pharmacologic Substance’, ‘Therapeutic or Preventive Procedure’
- Value: numeric or SNOMED adjective.

Abnormality Rule

- Attribute: UMLS semantic types: ‘Acquired Abnormality’, ‘Anatomical Abnormality’, ‘Congenital Abnormality’, ‘Mental or Behavioral Dysfunction’, ‘Pathologic Function’
- Value: none

Regional Abnormality Rule

- Attribute: UMLS semantic types: ‘Acquired Abnormality’, ‘Anatomical Abnormality’, ‘Cell or Molecular Dysfunction’, ‘Congenital Abnormality’, ‘Disease or Syndrome’, ‘Injury or Poisoning’, ‘Neoplastic Process’, ‘Pathologic Function’
- Value: UMLS semantic types: ‘Body Location or Region’, ‘Body Part, Organ or Organ’, ‘Component’, ‘Body Space or Junction’, ‘Body System’, ‘Embryonic Structure’, ‘Fully Formed Anatomical Structure’, ‘Tissue’, ‘Cell’, ‘Cell Component’, ‘Spacial Concept’

Modified Abnormality Rule

- Attribute: UMLS semantic types: ‘Acquired Abnormality’, ‘Anatomical Abnormality’, ‘Congenital Abnormality’, ‘Pathologic Function’
- Value: SNOMED adjective

Other changes to the FindX program included changes to the adjective list. First, additions and deletions were made to the SNOMED adjective list to account for values that were not well represented. Second, changes were made to account for inflectional variation of the adjectives. The SPECIALIST lexicon was used to convert the adjective list to its lexical base form. The FindX program was then modified to also convert the words in the input phrase to their lexical base and give the spelling variant as well. These were then compared to the base form of the adjective list. This accounts for confounders like the term “Faeces Discoloured” which is the British English spelling variant of Feces Discolored and which would not have been picked up by FindX without these modifications.

FindX was also modified to account for the ambiguity designators within the UMLS. For instance the term Blood Pressure <2> was falsely found to be a Finding by the FindX Physiologic Function Rule because it interpreted ‘<2>’ to be a valid numeric value. An initial Clean-up procedure stripped out the ambiguity designators to avoid this problem.

Results

Ontology for Clinical Findings

The ontology for clinical Findings was derived from the formal definition of Findings and the ontology principles for structuring roles and types. The Formal

Definition of Findings is that they are a kind of Information used as Evidence of the manifestation of Disease and the associated with healthcare Activities and an Attribute under study. Using the ontology principles of Guarino, Welty, and Puskejuvsky, we get the ontology as seen in Figure 2. In this representation, Findings are subsumed under the sortal type of Finding and the Role type of Evidence. Findings have an associative relationship to Disease. Information also has associative relationships to Activity and Attributes. However, these Activities and Attributes are placeholders for the Categories that have this type of associative relationship with Findings. The challenge is first to identify these associated categories as well as the subcategories for the Findings super type, then to populate these categories with UMLS concepts. To do this we turn to the results of the bottom up approach.

Bottom-up Results

Modifications to the FindX program resulted in improved Recall from a baseline of 62% to 75%. Precision remained relatively stable at 89%. Of the 94182 terms taken as Findings, 65798 or 70% were identified as Findings by the FindX program. In the original version of the FindX program, 85% of those terms identified as Findings triggered only the Finding Rule. The modified version of FindX decreased that number to 55% increasing the categorization of a considerable number of what were previously ambiguous findings.

Discussion

FindX Performance has improved. Continue work in populating the ontology where the bottom details meet the high level concepts.

Conclusion

We anticipate application to other data sets as well as the application to knowledge processing.

Figure 1 – Creation of Findings, Use of Findings, What is a finding?

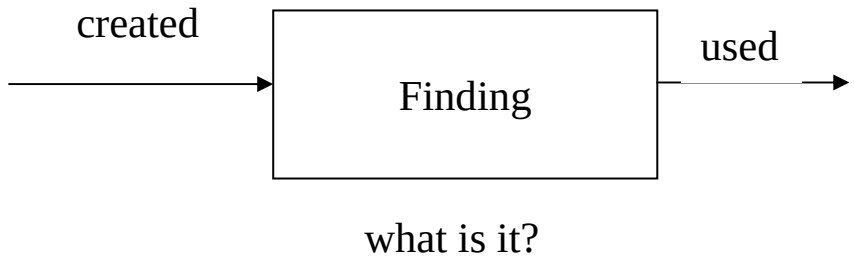
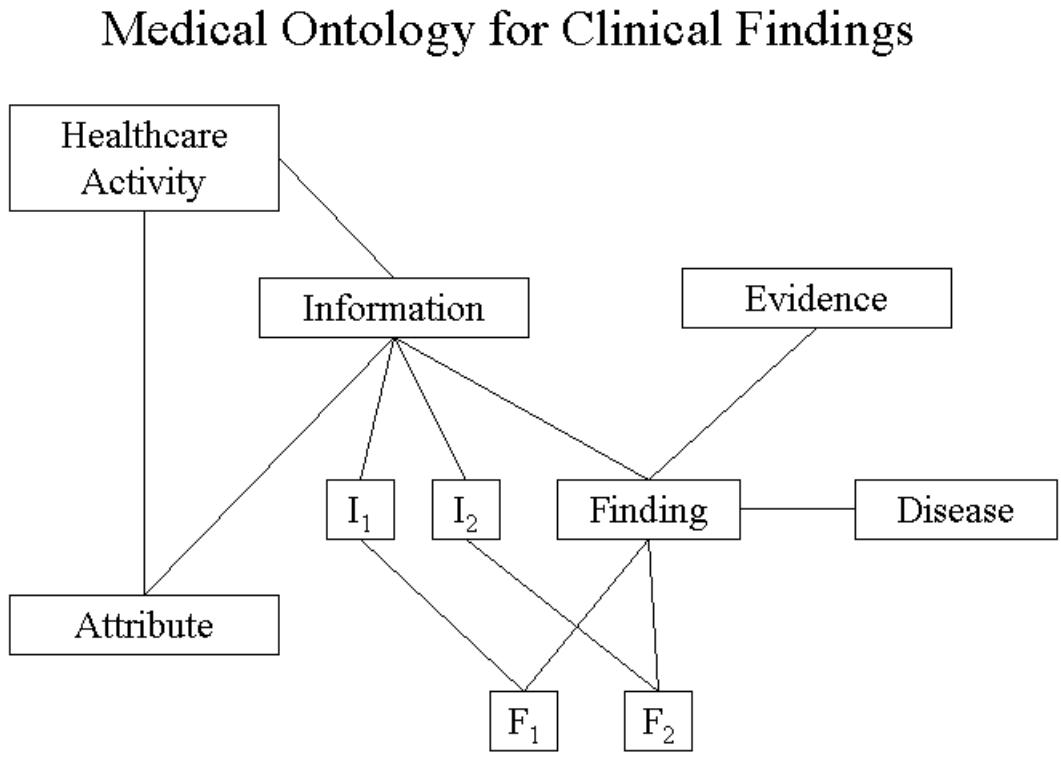


Figure 2: Ontology for Clinical Findings



Bibliography

Burgun A, Bodenreider O, Le Duff F, Mounssouni F, Loréal O. Representation of roles in biomedical ontologies: a case study in functional genomics. Proceedings of AMIA Annual Symposium 2002: (submitted).

Burnum, JF Medical Diagnosis through Semiotics: Giving meaning to the sign. Annals of Internal Medicine. 1993;119:939-943

Bodenreider O, Burgun A, Rindflesch TC. Assessing the consistency of a biomedical terminology through lexical knowledge. Proceedings of the Workshop on Natural Language Processing in Biomedical Applications (NLPBA'2002) 2002:(in press).

Guarino N, Welty C. A formal ontology of properties. Proceedings ECAI-2000, 2000.

Peirce CS. Logic as semiotic: the theory of signs. In: Buchler J; ed: Philosophical Writings of Peirce. New York: Dover Publications Inc.; 1955:98-119

Sneiderman Charles A.; Thomas C. Rindflesch; and Alan R. Aronson. 1996. Finding the Findings: Identification of Findings in Medical Literature Using Restricted Natural Language Processing. In Cimino JJ (ed.) Proceedings of the 1996 AMIA Annual Fall Symposium, 239-43.

Strawson PF. Individuals: an essay in descriptive metaphysics, Routledge, London, 1959
Sowa J.F. Conceptual Structure: Information Processing in mind and machine, Reading, MA: Addison Wesley, 1984

Sowa JF. Knowledge Representation. Brooks Cole, 2000.

Weed LL. Medical Records that Guide and Teach. N Engl J Med, 1968; 278(11): 593-600; 278(12): 652-7.

Handbook of Medical Informatics. JH van Bemmel, MA Musen

DeGowin's Diagnostic Examinatation.

Puskejsky, James. The generative Lexicon.